Medical Problem and Document Model for Natural Language Understanding

Stephane Meystre, MD, MS, Peter J. Haug, MD
Department of Medical Informatics, University of Utah, Salt Lake City, Utah

Abstract

We are developing tools to help maintain a complete, accurate and timely problem list within a general purpose Electronic Medical Record system. As a part of this project, we have designed a system to automatically retrieve medical problems from free-text documents. Here we describe an information model based on XML (eXtensible Markup Language) and compliant with the CDA (Clinical Document Architecture). This model is used to ease the exchange of clinical data between the Natural Language Understanding application that retrieves potential problems from narrative document, and the problem list management application.

Introduction

Growing interest in better ways of organizing the Electronic Medical Record has generated new attention to the medical Problem List, a classical part of the Problem-Oriented Medical Record. This paper presents a small part of a larger project, which consists in building an environment where the Problem List, is easily and effectively maintained. To this end, a Natural Language Understanding (NLU) application that will harvest potential Problem List entries from the multiple free-text electronic documents available in our EMR (Electronic Medical Record) will be developed and tested. These potential problems will drive an application designed for the management of the problem list in an environment where the computer does much of the initial problem identification. The problems identified will be proposed to the physicians for addition to the official Problem List.

To allow the exchange of clinical information between the NLU application and the problem list management application, an information model is needed and is described in this paper. The existence of such a common data model has been shown to be essential to knowledge exchange if laborious reworking is to be avoided[1]. In our case, it is used to represent medical problems and the documents from which they were extracted.

Background

More than three decades ago, Larry Weed proposed the problem-oriented medical record as a remedy for the complexity of the medical knowledge and clinical data, and for weaknesses in the documentation of medical care[2,3]. He noted the lack of consistent structure and content in the progress notes that make up a large part of the medical record. He proposed a standard approach emphasizing a list of patient problems that is scrupulously maintained by those caring for the patient. This problem list serves the dual purpose of providing a brief, formal summary of the patient's illnesses and of acting as a tool for organizing the routine documentation of the physician's decision-making process and the plan for and results of care.

The problem-oriented, Computer-based Patient Record (CPR) and the problem list have seen renewed interest as an organizational tool in the recent years[4-6], but most of today's patient records remain time-oriented.

The Institute of Medicine report on the CPR[7,8] recommends that it contain a problem list that specifies the patient's clinical problems and the status of each. It mentions advantages to this approach: the problem list can be the central place for clinicians to obtain a concise view of all patients problems; this list facilitates associating clinical information in the record to a specific problem; and the Problem List can encourage an orderly process of clinical problem solving and clinical judgment. The problem list in a problem–oriented patient record also provides a context in which continuity of care is supported, preventing both redundant and repeated actions[6].

At Intermountain Health Care (IHC) , a health maintenance organization serving Utah, a new version of our Clinical Information System (HELP 2) is in development. It features a problem-oriented medical record, and the problem list is therefore its central component. This problem list is maintained through web-based tools, and uses a terminology of about 60,000 concepts provided by the 3M Health Data Dictionary (HDD). Already in use in the outpatient setting, the current version of the problem list is often incomplete, inaccurate, out-of-date or even not used at all. The global aim of our project is to automate the process of creating and maintaining a problem list for hospitalized patients and thereby help to guarantee the timeliness, accuracy and completeness of this information.

The patient record contains a considerable amount of information, but, commonly, most of the recorded clinical information is unstructured text, also called free-text. These free-text documents largely represent patient history and reports of therapeutic interventions

or clinical progress and make up a substantial part of the medical record, providing information leading to the final diagnosis in 76% of the cases[9]. Free-text is still the most user-friendly and expressive way of recording information, but the increasing use of encoded data and the requirement for standard medical data set creates a need for coded information instead. As a possible answer to this problem, Natural Language Processing can convert narrative text into coded data, and therefore extend the use of the CPR[10].

Several groups have evaluated techniques for automatically encoding textual documents from the medical record. The Linguistic String Project has developed a series of tools for analyzing medical text[11]. X-ray reports appear to be an especially fertile ground for NLU. Two groups have developed systems whose focus is the radiologist's report of the chest x-ray. Zingmond has applied a semantic encoding tool to these reports to recognize abnormalities that should receive follow-up[12], and Friedman has studied techniques for encoding interpretations found in these reports[13,14]. In addition, Friedman and her colleagues have studied NLU in mammography reports[15], neuroradiology reports[16], and discharge summaries[17]. Good performance was demonstrated. Our Medical Informatics group at the LDS Hospital and the University of Utah has focused its NLU research on reports for chest radiographs[18-20] emphasizing pneumonia-related data[21,22]. The latest version of the NLU application, called MPLUS[23], provides a syntactic analysis based on a context-free grammar with a bottom-up chart parser, interleaved with the semantic analysis using Bayesian Networks (also called belief networks). This application is being adapted and developed to retrieve medical problems in many different types of free-text documents.

Information models for clinical data or documents facilitate the extraction of patient information, serve as a framework for combining patient data from multiple sources, and allow sharing medical decision-support logic and patient care applications. Some research has already been done in the development of medical data models, like the Event Model proposed by Huff, Rocha et al.[1], or the model proposed by Sager et al.[11] to facilitate document retrieval. Models for medical documents were also proposed, like the first ANSI-approved healthcare standard: the HL7 CDA (Clinical Document Architecture)[24]. It uses XML[25] to facilitate the exchange of documents between users. A successful prototyping of a CDA-based structured discharge summary system was implemented between the clinical environment and the community environment of family practice[26]. In addition, Friedman, Hripcsak et al.[27] proposed a document model designed using XML, and used an NLP (Natural Language Processing) application to automatically create an enriched structured document consistent with the model and containing references to identifiers in the original unstructured document.

XML is a data storage toolkit, a configurable vehicle for any kind of information, and an evolving open standard embraced by everyone. It can store and organize any kind of data, offers many ways to check the quality of documents, and is easy to read and parse by humans and programs alike. It is a subset of SGML (Standard Generalized Markup Language, ISO standard 8879:1986)[28], and like the latter, it is not itself a markup language like HTML, but a set of rules for building markup languages. XML documents are validated against DTDs (Document Type Definition) or XML Schemata (also called XSD)[25]. The latter is the new version of metadata for XML documents. Like the DTDs, XML Schemata provide a mean for defining the structure, content and semantics of XML documents, but have advantages like being themselves written in XML and providing better data types definition. Even if not yet part of the official XML specification, they will probably soon replace the DTDs. In the healthcare field, many authors report the use of SGML or XML to tag medical documents[29-32].

The Unified Modeling Language (UML)[33] is a graphical notation used to express software and data designs, and is the successor of the wave of object-oriented analysis and design methods. It is widely adopted as the de facto industry standard and is an OMG (Object Management Group) standard. In the healthcare domain, some authors recently reported the use of UML for analysis and design of secure HIS (Health Information Systems)[34], for HIS architecture description[35], and for system design in public health informatics[36].

Model Description

The development process for the models described here began with the analysis of the processing steps of the NLU application and of the problem list management application. Research to identify the desired characteristics for problems in the list followed. The final selection comprises twenty characteristics, inserted in the medical problem model.

Two related models were developed to allow linking of the problem to the document and to the sentence(s) the problem was extracted from: the Medical Problem Model and the Medical Document Model. This will give users the ability to track the source of the proposed problems, as shown on figure 1. This back

link is also an essential feature to enable traceability and quality control of the NLU results. These two models form the Information Model of our system.

The Terminology Model is based on the HDD cited above. This model uses a subset of a *Problem* hierarchy that includes one hundred problems of diagnosis type (as opposed to problems of history, investigation, or sign/symptom type). Following development and prototyping, the model will be extended to all the problems present in the HDD. The hundred problems were selected based on their frequency of use at IHC in general and in the specialized domain of the future implementation of our project (cardiovascular).

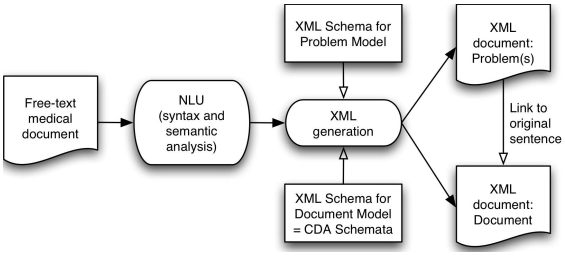

Fig. 1: Diagram of the system

The models were conceived and represented in UML, as depicted in figure 2, and implemented in XML, as XML Schemata. XML was selected for its increasing use and for the availability of numerous commercial and publicly available software.

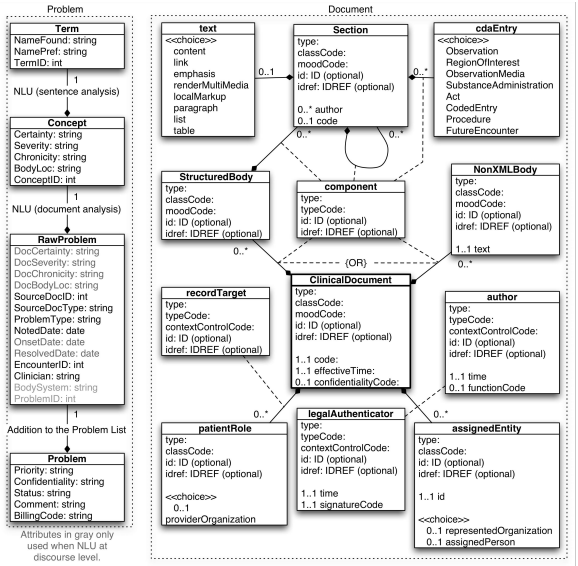

Fig. 2: Medical Problem and Document Model (UML class diagram, simplified from the CDA RIM (Reference Information Model)).

The XML Schemata will be used by the application generating the XML output, to ensure validity of the latter (see figure 1). The two following sections describe these two models in more details.

Medical Problem Model:

In this model, the problems were split into four different components, corresponding to each main level of free-text document processing by the NLU application and subsequent management by the Problem List user. The *Term* is at the word level, the *Concept* at the sentence level, the *RawProblem* at the document level, and the *Problem* at the problem list level, after addition and refinement by the user of the list. Components are related by composition relationships ("is a"), meaning that one *Term* is related to one *Concept(s)*, that one *Concept* is related to one *RawProblem(s)*, etc.

Medical Document Model:

Among the necessary features of a problem list, as suggested by Campbell[4], a history of change is cited. All transactions on items in the problem list will be recorded. To allow users of the list to track the source of the proposed problem at the sentence level, we first designed a simple document model composed of three elements (*Document, Section(s),* and *Sentence(s)*). The result would have been a model giving the source tracking capability, but only compatible with our system. To take full advantage of the NLU and XML generation processes, we then decided to base our Document Model on an existing standard: HL7's CDA. This decision made the Model more complex, but will allow better compatibility and exchange of the resulting XML documents, since these will be CDA-compliant. Our development is based on the latest version of the CDA: the Release Two (April 2003), rather than the Release One, even if the latter is already a recognized standard and the former only a working draft.

Relationships between the models will be multiple, relating Problem characteristics to some CDA classes or attributes, like *Clinician* to *ClinicalDocument.author*, *Source* to *ClinicalDocument.code*, *ConceptID* to *StructuredBody...Observation.code*, etc.

Validation through an example:

To demonstrate the use of these models, the following example is proposed, from the free-text note (figure 3) to the XML documents (figures 4 and 5). In this case, the problem retrieved by the NLU application would be "acute abdominal pain", beginning with "pain" at the *Term* level, and adding the chronicity "acute" and anatomical location "abdomen" at the *Concept* level. Further analysis of the note at the discourse and document level would give the information needed to instantiate the models.

CONSULTATION NOTE:
Consultant: Stephane Meystre, MD
Patient: xxx Date: August 11th 2002
History of Present Illness:
Patient is a 25-year-old male, referred for acute abdominal pain. Onset of the pain one day ago, in the periumbilical region, with subsequent migration to the right iliac fossa…

Fig.3: Consultation note.

Part of the resulting XML document for the note is shown in figure 4:

```
<!--CDA header –>
 <id extension="12345678" root="IHC"/>
 <code code="11488-4" codeSystem="LOINC"
    displayName="CONSULTATION NOTE"/>
 <effectiveTime value="20020811"/>
 <author>
    …   <name>
           <given>Stephane</given>
           <family>Meystre</family>
           <suffix>MD</suffix>
        </name>   …
 </author>
 <recordTarget>
    <patientRole>
       …
    </patientRole>
 </recordTarget>
<!--CDA body-->
 <component>
    <StructuredBody>
       <component>
          <section id="s1">
             <code code="10164-2" codeSystem="LOINC"
                displayName="History of Present Illness"/>
             <text>
                Patient is a 25 year old male, referred for
                <content ID="p1">acute abdominal pain
                </content>. Onset of the pain one day ago,
                in the periumbilical region, with subsequent
                migration to the right iliac fossa.
             </text>
             <component1>
                <Observation>
                   <code code="1450279" codeSystem=
                   "IHC_NCID" displayName="abdominal
                      pain"/>
                   <effectiveTime value="20020811"/>
                   <reference idref="p1"/>
                </Observation>
                …
```

Fig. 4: XML document for the consultation note.

The resulting XML document for the problem is shown in figure 5:

```
<?xml version="1.0" encoding="UTF-8"?>
<Problem xmlns:xsi="http://www.w3.org/…/ProbModel.xsd">
<Priority>medium</Priority>
<Confidentiality>confidential</Confidentiality>
<Status>active</Status>
<RawProblem>
   <DocCertainty>definite</DocCertainty>
   <DocChronicity>acute</DocChronicity>
   <DocBodyLoc>abdomen</DocBodyLoc>
   <Source>123456</Source>
   <ProblemType>symptom</ProblemType>
   <NotedDate>2002-08-11</NotedDate>
   <OnsetDate>one day ago</OnsetDate>
   <EncounterID>789000</EncounterID>
   <Clinician>Stephane Meystre, MD</Clinician>
   <BodySystem></BodySystem>
   <ProblemID>1450279</ProblemID>
   <Concept>
      <Certainty>definite</Certainty>
      <Chronicity>acute</Chronicity>
      <BodyLoc>abdomen</BodyLoc>
      <ConceptID>1450279</ConceptID>
      <Term>
         <NameFound>pain</NameFound>
         <NamePref>Pain</NamePref>
         <TermID>30722</TermID>
      </Term>
   </Concept>
</RawProblem>
</Problem>
```

Fig. 5: XML document for "acute abdominal pain".

Discussion

The purpose of the models described in this paper is to ease the exchange of clinical data between different clinical applications and subsystems. These models accommodate the representation of medical problems and their expression in clinical documents, with a link between the retrieved problem and its source at the word level. The current Problem List at IHC and the enhanced version we are developing are and will be based on pre-coordinated concepts, avoiding composition that adds a lot of complexity and is barely used and prone to errors.

To be able to improve the sensitivity and precision of our NLU application, *unknown* concepts will be tagged. This will later be used to complement our knowledge base and refine the NLU application.

The planned future work on this project will consist of the NLU application development for text analysis at the discourse and document level. We will make changes to the underlying technologies to improve the sensitivity and precision of the tool. We will also extend our system to "simple" problems (e.g. findings). The needed knowledge base for finding-disease, finding-finding, and disease-disease relationships will be based on the HDD ontology, eventually complemented with knowledge bases already used for this purpose, like QMR[5] (Quick Medical Reference). The NLU application will then be evaluated in a laboratory resource's function study, for recall (sensitivity) and precision (positive predictive value). The last step will be development and implementation of the problem list management application in an inpatient setting (cardiovascular ward). The evaluation will be a field resource's function study.

References

1. Huff SM, Rocha RA, Bray BE, Warner HR, Haug PJ. An event model of medical information representation. J Am Med Inform Assoc 1995;2(2):116-34.
2. Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(11):593-600.
3. Weed LL. Medical records that guide and teach. N Engl J Med 1968;278(12):652-7 concl.
4. Campbell JR. Strategies for problem list implementation in a complex clinical enterprise. Proc AMIA Symp 1998:285-9.
5. Starmer J, Miller R, Brown S. Development of a Structured Problem List Management System at Vanderbilt. Proc AMIA Annu Fall Symp 1998:1083.
6. Bayegan E, Tu S. The helpful patient record system: problem oriented and knowledge based. Proc AMIA Symp 2002:36-40.
7. Institute of Medicine (U.S.). Committee on Improving the Patient Record, Dick RS, Steen EB, Detmer DE. The computer-based patient record : an essential technology for health care. Rev. ed. Washington, D.C.: National Academy Press; 1997.
8. Warren JJ, Collins J, Sorrentino C, Campbell JR. Just-in-time coding of the problem list in a clinical environment. Proc AMIA Symp 1998:280-4.
9. Peterson MC, Holbrook JH, Von Hales D, Smith NL, Staker LV. Contributions of the history, physical examination, and laboratory investigation in making medical diagnoses. West J Med 1992;156(2):163-5.
10. Spyns P. Natural language processing in medicine: an overview. Methods Inf Med 1996;35(4-5):285-301.
11. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1994;1(2):142-60.
12. Zingmond D, Lenert LA. Monitoring free-text data using medical language processing. Comput Biomed Res 1993;26(5):467-81.
13. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc 1994;1(2):161-74.
14. Hripcsak G, Kuperman GJ, Friedman C. Extracting findings from narrative reports: software transferability and sources of physician disagreement. Methods Inf Med 1998;37(1):1-7.
15. Jain NL, Friedman C. Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. Proc AMIA Annu Fall Symp 1997:829-33.
16. Elkins JS, Friedman C, Boden-Albala B, Sacco RL, Hripcsak G. Coding neuroradiology reports for the Northern Manhattan Stroke Study: a comparison of natural language processing and manual review. Comput Biomed Res 2000;33(1):1-10.
17. Friedman C, Knirsch C, Shagina L, Hripcsak G. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. Proc AMIA Symp 1999:256-60.
18. Haug PJ, Ranum DL, Frederick PR. Computerized extraction of coded findings from free-text radiologic reports. Radiology 1990;174(2):543-8.
19. Haug P, Koehler S, Lau LM, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. Proc Annu Symp Comput Appl Med Care 1994:247-51.
20. Haug PJ, Koehler S, Lau LM, Wang P, Rocha R, Huff SM. Experience with a mixed semantic/syntactic parser. Proc Annu Symp Comput Appl Med Care 1995:284-8.
21. Fiszman M, Chapman WW, Evans SR, Haug PJ. Automatic identification of pneumonia related concepts on chest x-ray reports. Proc AMIA Symp 1999:67-71.
22. Fiszman M, Chapman WW, Aronsky D, Evans RS, Haug PJ. Automatic detection of acute bacterial pneumonia from chest X-ray reports. J Am Med Inform Assoc 2000;7(6):593-604.
23. Christensen L, Haug P, Fiszman M. MPLUS: a probabilistic medical language understanding system. Proceedings of the Workshop on Natural Language Processing in the Biomedical Domain 2002:29-36.
24. Dolin RH, Alschuler L, Beebe C, Biron PV, Boyer SL, Essin D, et al. The HL7 Clinical Document Architecture.J Am Med Inform Assoc 2001;8(6):552-69
25. Bray T, Paoli J, Sperberg-McQueen C, Maler E. Extensible Markup Language (XML) 1.0 (Second Edition) 2000. http://www.w3.org/TR/REC-xml
26. Paterson G, Shepherd M, Wang X, Watters C, Zitner D. Using the XML-based Clinical Document Architecture for Exchange of Structured Discharge Summaries. Proc. 35th Hawaii Int Conf on System Sciences 2002.
27. Friedman C, Hripcsak G, Shagina L, Liu H. Representing information in patient reports using natural language processing and the extensible markup language. J Am Med Inform Assoc 1999;6(1):76-87.
28. Bingham H. SGML Syntax Summary Table of Contents. In; 1996.
29. Sager N, Nhan NT, Lyman M, Tick LJ. Medical language processing with SGML display. Proc AMIA Annu Fall Symp 1996:547-51.
30. Dolin RH, Rishel W, Biron PV, Spinosa J, Mattison JE. SGML and XML as interchange formats for HL7 messages. Proc AMIA Symp 1998:720-4.
31. Zweigenbaum P, Bouaud J, Bachimont B, Charlet J, Seroussi B, Boisvieux JF. From text to knowledge: a unifying document-centered view of analyzed medical language. Methods Inf Med 1998;37(4-5):384-93.
32. Krauthammer M, Hripcsak G. A knowledge model for the interpretation and visualization of NLP-parsed discharged summaries. Proc AMIA Symp 2001:339-43.
33. OMG. Unified Modeling Language (UML), v.1.4. http://www.omg.org/uml/
34. Blobel B, Roger-France F. A systematic approach for analysis and design of secure health information systems. Int J Med Inf 2001;62(1):51-78.
35. Winter A, Brigl B, Wendt T. A UML-based ontology for describing hospital information system architectures. Medinfo 2001;10(Pt 1):778-82.
36. Orlova A, Lehmann H. A UML-based Meta-Framework for System Design in Public Health Informatics. Proc AMIA Symp 2002:582-6.